Fundamental Issues in Biometric Performance Testing: A modern statistical and philosophical framework for uncertainty assessment

James L. Wayman, SJSU
Antonio Possolo, NIST
Tony Mansfield, NPL

Historical Motivation

San José State UNIVERSITY

- Response to the Biometric Consortium's 1994 questions to the community on repeatability and reproducibility of test results
  - Why does field technical performance of biometric systems differ from laboratory performance (often by orders of magnitude)?
  - How should "confidence intervals" be calculated on laboratory test results?

Fundamental Position

San José State UNIVERSITY

- A test of biometric performance must inherit the framework of scientific testing in general
- Uncertainty is one component impacting repeatability and reproducability
- Uncertainty in biometric performance measures can be discussed within the current NIST and ISO framework of measurement uncertainty assessment.
- The biometrics community must constructively address the issues of repeatability, reproduceability and performance prediction within the received framework of scientific testing.

Forms of Biometric Testing

San José State UNIVERSITY

- Useability
- Vulnerability
- Standards compliance
- Reliability/Availability/Maintainability
- Return-on-Investment
- Technical (of lesser practical importance?)
  - Error rates
  - Throughput

Our Key Remarks

1. "Uncertainty" is a broader concept than "error", as historically understood; it is the doubt about how well the test result represents the quantity measured (or being said to be measured). Uncertainty can exist even in the absence of error.
2. A central source of uncertainty is definitional incompleteness in specifying the "unit of empirical significance" for the measurand – full specification of which would require a "infinite amount of information".
3. What we are measuring is often only a proxy for the measurand of real interest, even if fully defined, which adds yet another source of uncertainty in our measurement.
4. How we control, measure and report the values in a test must reflect how we expect those values to be used by others. In other words, our testing and reporting must take into account, and state, how we expect the results to be used.

Duhem-Quine Thesis and Testing Holism

- "…the physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses"" – Duhem, 1906
- the results of any scientific test reflect the totality of conditions of the test ("the unit of empirical significance"), including instrumentation, background assumptions, auxiliary hypotheses, and even the theories being tested themselves. So what we measure in any experiment is the totality of all the elements existing in both the physical and intellectual environment of the test and, further, the measurements must be expressed using words and concepts that themselves may be subject to change as our understanding progresses.

The NIST Tradition

- 1947 founding of SEL (now SED) with Churchill Eisenhart as Director.
- "In these days when so much emphasis is properly being placed on economy of government research operations, it is important to take advantage of the substantial savings which can be effected by substituting sound mathematical analysis for costly experimentation. In science as well as in business, it pays to stop and figure things out in advance" -- NBS Director Edward Condon
- Repeatability and reproducibility as key goals.
- "a measurement operation must have attained …a state of statistical control . . . before it can be regarded in any logical sense as measuring anything at all" – Churchill Eisenhart

"Chance Causes" and "Common Causes"

- Both limit the repeatability and reproducability of results
- Measurements require an associated statement of uncertainty stemming from both chance causes and common causes, "and their relative importance in relation to the intended use of the reported value, as well as to other possible uses to which it may be put" -- Eisenhart

Statistical Control of the Unit of Empirical Significance

An example:

- "What is the speed of sound?"
  - What medium?
    - In air
      - At what precision do we need this measure
        - » High?
          - » What temperature?
          - » What pressure?
          - » What molecular composition?
          - » Unknown factors within the unit of empirical significance (humidity, salt content, moon phase…)

International Committee on Weights and Measures (CIPM), 1980

"The uncertainty in …a measurement generally consists of several components which may be grouped into two categories according to the way in which their numerical value is estimated:

A. those which are evaluated by statistical methods,

B. those which are evaluated by other means.

There is not always a simple correspondence between the classification into categories A or B and the previously used classification into "random" and "systematic" uncertainties. The term "systematic uncertainty" can be misleading and should be avoided.

Any detailed report of the uncertainty should consist of a complete list of the components, specifying for each the method used to obtain its numerical value "-- Bureau International des Poids et Mesures (BIPM) Recommendation INC-1, as adopted by CIPM

Type A and Type B Estimations

- Type A: Evaluation by statistical methods means estimation of a component of uncertainty using statistical methods applied to **replicated indications** obtained during measurement.
- Type B: Other means of evaluation include information derived from authoritative publications, for example in the certificate of a certified reference material, or based on expert opinion.
- Appears to combine frequentist and subjective measures in a way that neither frequentists nor Bayesians can endorse
- GUM revision underway in Working Group 1 of the Joint Committee for Guides in Metrology (BIPM)

NIST Response

- "… this NIST policy adopts in substance the approach to expressing measurement uncertainty recommended (CIPM)" -- "Statements of Uncertainty Associated With Measurement Results," Appendix E, NIST Technical Communications Program, Subchapter 4.09 of the Administrative Manual
- Recommends combined uncertainty be multiplied by 2 to create "expanded uncertainty".
- B. Taylor and C. Kuyatt, "Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results", NIST Technical Note 1297, 1994 Edition

ISO Guide 98, "Guide to Expression of Uncertainty in Measurement"

- "… in principle, a measurand cannot be *completely* described without an infinite amount of information. Thus, to the extent that it leaves room for interpretation, incomplete definition of the measurand introduces into the uncertainty of the result of a measurement a component of uncertainty that may or may not be significant relative to the accuracy required of the measurement'"
- "..when all of the known or suspected components of error have been evaluated and the appropriate corrections have been applied, there still remains an uncertainty about the correctness of the stated result, that is, a doubt about how well the result of the measurement represents the value of the quantity being measured"

The ISO Concept of Uncertainty

	Type A Statistical	Type B Other
Random	Classic "confidence intervals"	
Systematic		

Which type of error (random, systematic) dominates and how should it be estimated?

Type B estimation allows for expert opinion.

Neyman's "Confidence Intervals"

- Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, Vol. 236, No. 767 (Aug. 30, 1937), pp. 333-380

X—Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability

By J. Neyman

Reader in Statistics, University College, London

(Communicated by H. Jeffreys, F.R.S.—Received 20 November, 1936—Read 17 June, 1937)

Contents

I—Introductory

(*a*) *General Remarks, Notation, and Definitions*

We shall distinguish two aspects of the problems of estimation : (i) the practical and (ii) the theoretical. The practical aspect may be described as follows :

(*ia*) The statistician is concerned with a population, π, which for some reason or other cannot be studied exhaustively. It is only possible to draw a sample from this population which may be studied in detail and used to form an opinion as to the values of certain constants describing the properties of the population π. For example, it may be desired to calculate approximately the mean of a certain character possessed by the individuals forming the population π, etc.

(*ib*) Alternatively, the statistician may be concerned with certain experiments which, if repeated under apparently identical conditions, yield varying results. Such experiments are called random experiments, (*see* p. 338). To explain or describe

VOL. CCXXXVI.—A 767 (Price 6s.) 2 Z [Published August 30, 1937

Neyman's Applications of "Confidence Intervals"

San José State UNIVERSITY

- "(ia) The statistician is concerned with a population, π, which for some reason or other cannot be studied exhaustively. It is only possible to draw a sample from this population which may be studied in detail and used to form an opinion as to the values of certain constants describing the properties of the population ,π…..
- (ib) Alternatively, the statistician may be concerned with certain experiments which, if repeated under apparently identical conditions, yield varying results."

A Different Approach by GUM

- Subsumes Neyman “confidence intervals” , but covers a much broader range of conditions, including experiments which cannot be repeated under identical conditions, as in biometrics
- “interval”: possible values of the measurand given combined random/systematic uncertainty evaluated by Type A and Type B methods
- “level of confidence” to describe the estimated probability that the measurand lies within that interval

Application to Biometric Performance Tests

- Technology
- Scenario
- Operational
  - Philips, Martin, Wilson, and Pryzbocki (2000)

Technology Tests

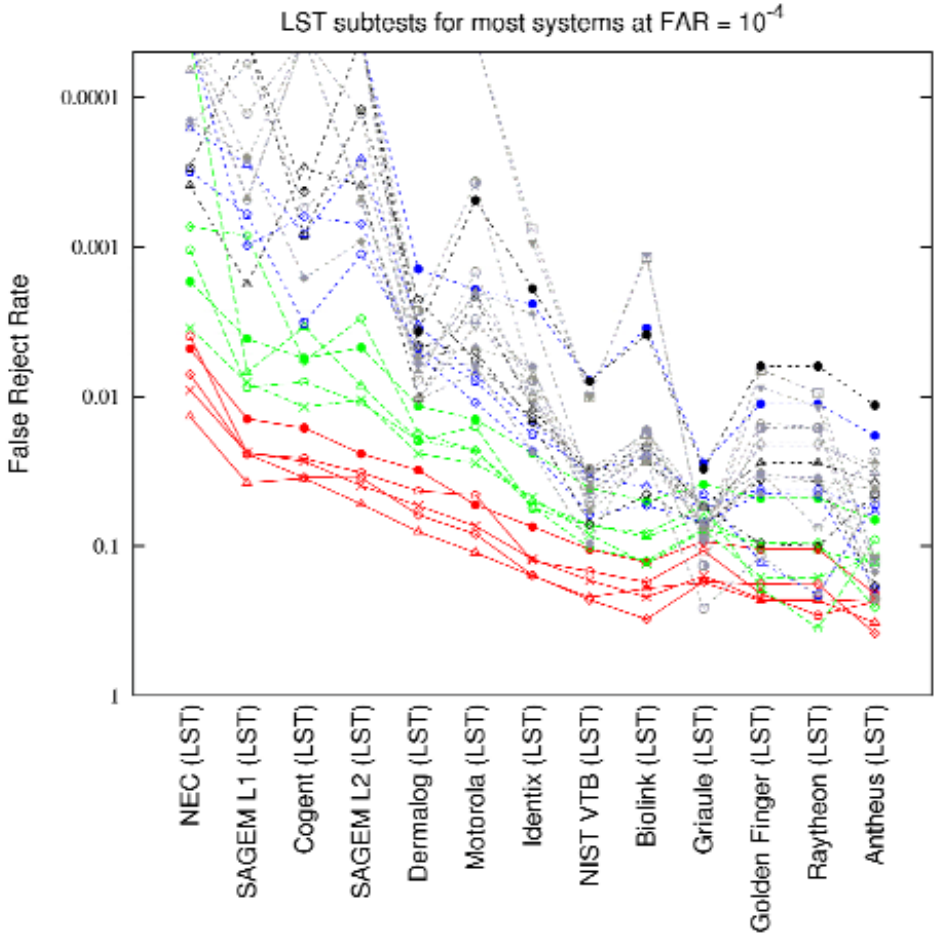
- Model: NIST "Proprietary Fingerprint Template Testing"

	DHS2	DOS	POE	POEBVA
D	0.9917	0.9845	0.9955	0.9932
F	0.9893	0.9944	0.9979	0.9979
H	0.9870	0.9978	0.9993	0.9994
I	0.9904	0.9978	0.9992	0.9992

- Measurand: TAR=(1-FNMR) at FMR=0.0001 for algorithm X against database Y
- Completely repeatable and reproduceable within hardware truncation limits
- Systematic uncertainty: Actual measurand (error rate against test key) is a proxy for stated measurand
- No "confidence intervals"because nothing is repeated under identical conditions and no data is a random sample of any larger population

IAD Approach to Matching Key Error Uncertainty

- FpVTE, NISTIR 7123



Scenario Tests

- Measurands in technology test
  - False match/non-match rates (FMR/FNMR)
- Measurands in (access control) scenario
  - False accept/false reject rates (FAR/FRR)
- Technology test results translate only with systematic uncertainty to scenario tests:
  - Uncertainty in models for FAR/FRR = F(FMR, FNMR)
  - Differences in algorithms, collection and data subject conditions
    - i.e: height and angle of acquisition device, form of instruction, influence of threshold, acoustic noise
- Repeatability and reproducibility of scenario tests only within control of all (including unknown) influencing factors.
- No random sampling or repetition under identical conditions

Operational Tests

- Measurands
  - System rejection/system acceptance
- Uncertainty in System False Acceptance
  - No independent assessment of ground truth after system acceptance
- Systematic uncertainty in System False Rejection
  - System rejections can result from non-biometric decisions
- Lack of repeatability and reproducability
  - All previous systematic uncertainties
  - Lack of ground truth for acceptance
  - Incommensurability of system rejection with biometric rejection

Concluding Remarks

1. "Uncertainty" is a broader concept than "error", as historically understood; it is the doubt about how well the test result represents the quantity measured (or being said to be measured). Uncertainty can exist even in the absence of error.
2. A central source of uncertainty is definitional incompleteness in specifying the "unit of empirical significance" for the measurand – full specification of which would require a "infinite amount of information".
3. What we are measuring is often only a proxy for the measurand of real interest, even if fully defined, which adds yet another source of uncertainty in our measurement.
4. How we control, measure and report the values in a test must reflect how we expect those values to be used by others. In other words, our testing and reporting must take into account, and state, how we expect the results to be used.

Recommendations

- We must expand, through controlled testing, our understanding of variables of influence within our "unit of empirical significance" in a biometric **system**.
- We must define our measurands more carefully.
- If uncertainty estimation is required, biometric testing and reporting should adopt the approach of GUM and NIST policy
  - All known sources of uncertainty (both random and systematic) should be reported.
  - Unknown sources of uncertainty should be expected.
  - Bounds should be established with expression of experimenter confidence that measureand lies within them.